COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

CENTRALE NANTES

Titre : Tests à noyaux, et leurs applications aux données de séquençage en cellule unique

Mot clés : tests à noyaux, séquençage en cellule unique, modèle linéaire à noyaux, Nyström, fonctions d'influence

Résumé : Les technologies de sequençage en cellule unique mesurent des informations à l'échelle de chaque cellule d'une population. Les données issues de ces technologies présentent de nombreux défis : beaucoup d'observations en grande dimension et souvent parcimonieuses. De nombreuses expériences de biologie consistent à comparer des conditions. L'objet de la thèse est de développer un ensemble d'outils qui compare des échantillons de données issues des technologies de séquençage en cellule unique afin de détecter et décrire les différences qui existent. Pour cela, nous proposons d'appliquer les tests de comparaison de deux échantillons basés sur les méthodes à noyaux existants. Nous proposons de généraliser ces tests à noyaux pour les designs expérimentaux quelconques, ce test s'inspire du test de la trace de Hotelling-Lawley. Nous implémentons pour la première fois ces tests à noyaux dans un package R et Python nommé *ktest*, et nos applications sur données simulées et issues d'expériences démontrent leurs performances. L'application de ces méthodes à des données expérimentales permet d'identifier les observations qui expliquent les différences détectées. Enfin, nous proposons une implémentation efficace de ces tests basée sur des factorisations matricielles de type Nyström, ainsi qu'un ensemble d'outils de diagnostic et d'interprétation des résultats pour rendre ces méthodes accessibles et compréhensibles par des non-spécialistes.

Title: Kernel-based testing and their applications to single-cell data

Keywords: kernel testing, single-cell, kernel linear model, Nyström, influence functions

Abstract: Single-cell technologies generate data at the single-cell level. They are coumposed of hundreds to thousands of observations (i.e. cells) and tens of thousands of variables (i.e. genes). New methodological challenges arose to fully exploit the potentialities of these complex data. A major statistical challenge is to distinguish biological information from technical noise in order to compare conditions or tissues. This thesis explores the application of kernel testing on single-cell datasets in order to detect and describe the potential differences between compared conditions. To overcome the limitations of existing kernel two-sample tests, we propose a kernel test inspired from the Hotelling-Lawley test that can apply to any experimental design. We implemented these tests in a R and Python package called *ktest* that is their first user-oriented implementation. We demonstrate the performances of kernel testing on simulated datasets and on various experimental single-cell datasets. The geometrical interpretations of these methods allows to identify the observations leading a detected difference. Finally, we propose a Nyström-based efficient implementation of these kernel tests as well as a range of diagnostic and interpretation tools.